Regresión y Data Mining

Silvia N. Pérez

Buscando respuestas…

- Se quiere predecir las ventas de juguetes. Se dispone de información de gastos de publicidad, población, etc.
- Se quiere determinar el impacto del gasto en publicidad en las ventas de juguetes.
- Se quiere evaluar el riesgo de enfermar de neumonía a partir de datos de salud de pacientes.
- Se quiere clasificar mensajes en Spam/NoSpam a partir de datos observados de este tipo de mensajes.

Data Mining trata de responder a partir de algoritmos para..

- *Predicción* – puede mostrar cómo ciertos atributos se comportarán en el futuro.
- *Clasificación* – puede separar los datos de manera que las diferentes clases o categorías pueden ser identificados sobre la base de combinaciones de parámetros.
- *Optimización* – Un objetivo final de la minería de datos puede ser la de optimizar el uso de recursos limitados, tales como tiempo, espacio, dinero o materiales y maximizar outputs, tales como ventas o beneficios bajo un determinado conjunto de restricciones.

Regresión y data mining

Una tarea de DM es tratar de **explicar** y **predecir** el comportamiento de alguna/s variable de interés analizando datos disponibles. Estos son justamente objetivos del análisis de regresión.

Un modelo de regresión asume que la media de esta variable, condicional a las variables inputs, es función (lineal?) de estas.

Porqué usar modelos de regresión?

- Son sencillos y permiten una descripción fácilmente interpretable de cómo los inputs afectan el output.
- Es bien conocida la inferencia en estos modelos.
- Dan buenos diagnósticos a partir de análisis de residuales, aún para grandes conjuntos de datos.
- Para predicción pueden superar modelos más sofisticados.
- Son la base de otros modelos (otras bases funcionales, modelos lineales generalizados, etc.)

Algunos modelos

$Y \sim X_1, \ldots$

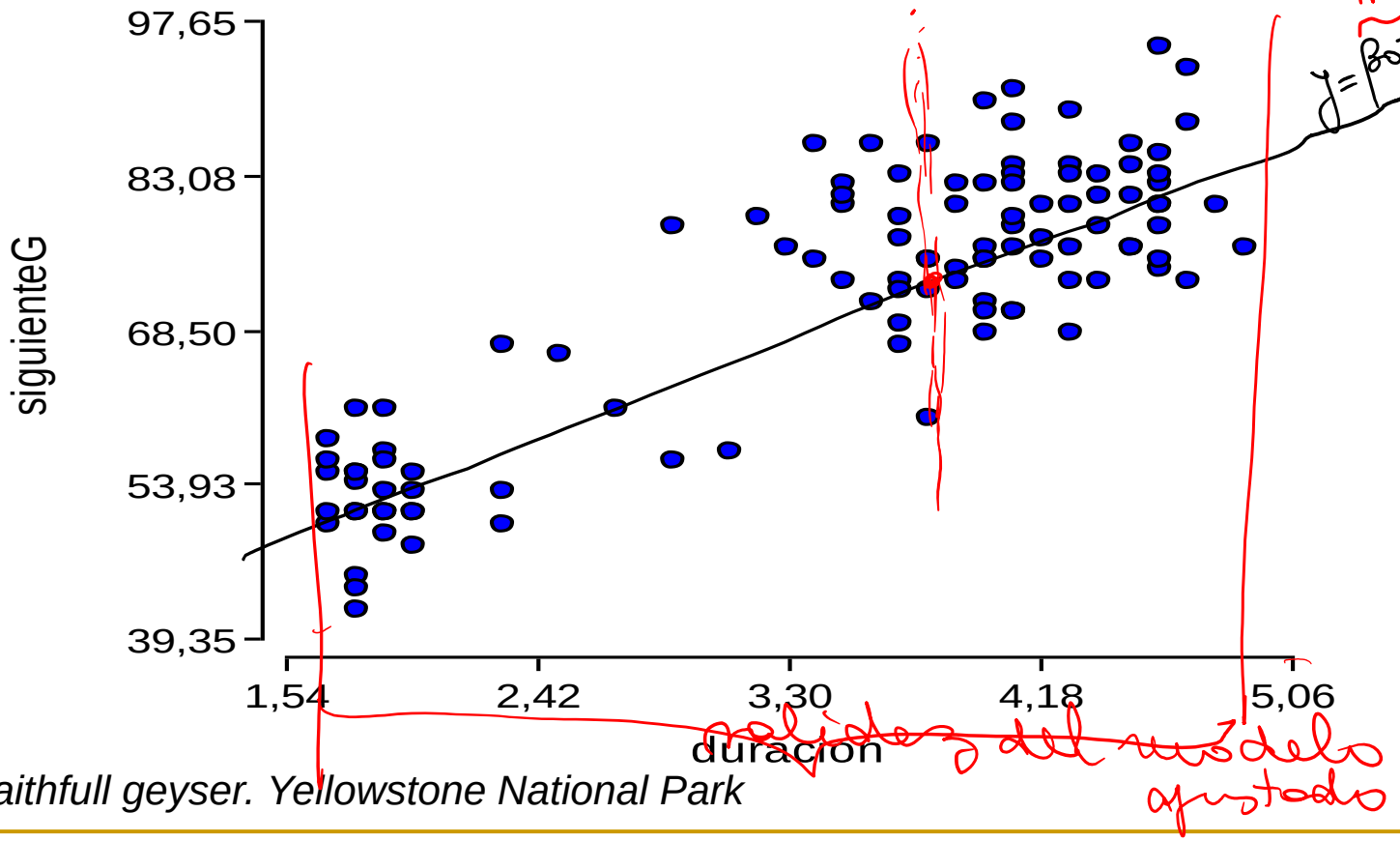
- Output condicional a las variables input es Normal con desvío constante: Modelos de Regresión
- Output binaria (Bernoulli): Regresión Logistica
- Datos de conteo (número de eventos en longitud/tiempo): Output con distribución *Poisson:* Regresión Poisson
- Output continua, regresores categóricos: Anova
- Etc…

Regresión Lineal Simple

Bibliografía:

- Chaterjee, S.; Hadi, A.; Price, B. "Regression Analysis by Example". Wiley (Cap. 2)
- Montgomery, Peck y Vining. "Introducción al Análisis de Regresión Lineal". (Cap. 2)
- Draper, N.R.; Smith H. "Applied Regression Analysis". 2nd ed. Wiley N.Y.

¿Puede predecirse la siguiente erupción de un geyser?



Old Faithfull geyser. Yellowstone National Park

Old Faithfull geyser

Los datos muestran que, aunque las erupciones no se producen a intervalos regulares, hay cierta relación entre la duración de una erupción y el tiempo que transcurre hasta la siguiente.

Los guardas han observado que cuanto más larga es una erupción, más tarda en presentarse la siguiente.

Se quiere llevar a turistas a ver el geyser en el momento oportuno. ¿Puede predecirse la próxima erupción?

Para qué sirve el análisis de regresión?

a) Describir/modelar la relación entre X e Y

b) Predecir valores de Y a partir de X

c) Probar hipótesis acerca de los parámetros del modelo (i.e.: cómo afecta un cambio en X a la Y)

El modelo de Regresión Lineal Simple

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

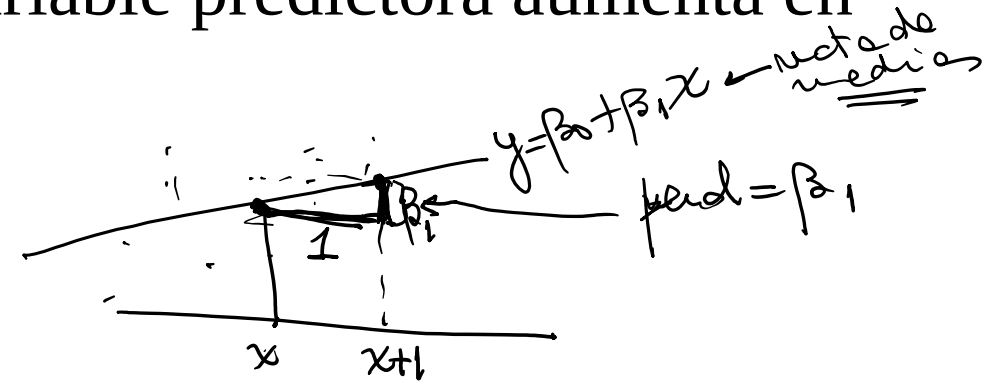
Considerando observaciones (x_i, y_i) para i=1,…n

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Donde ε_i es un error aleatorio con media **0** y varianza σ^2

Interpretación de los coeficientes de regresión

La pendiente indica el cambio promedio en la variable de respuesta cuando la variable predictora aumenta en una unidad.

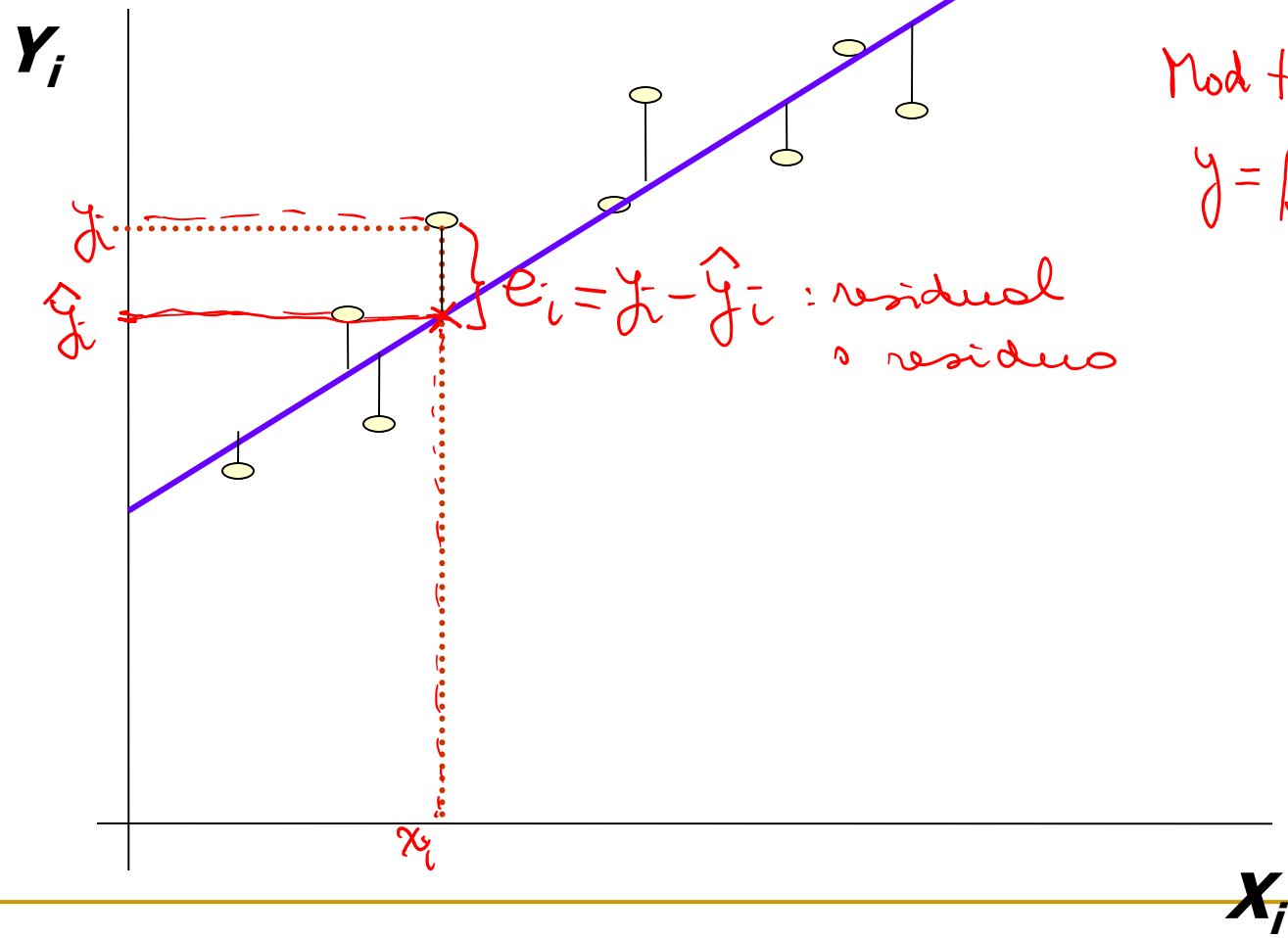


El **intercepto** indica el valor promedio de la variable de respuesta cuando la variable predictora vale 0.

Estimación de la línea de regresión usando Mínimos Cuadrados

recta ajustada o estimada

$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$



Mod teórico:

$y = \beta_0 + \beta_1 X + \varepsilon$

recta teórica

Estimación de la línea de regresión usando Mínimos Cuadrados

Se debe minimizar $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$

$(y_i - \hat{y}_i)^2$

de donde resulta que

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

Se deduce también que $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

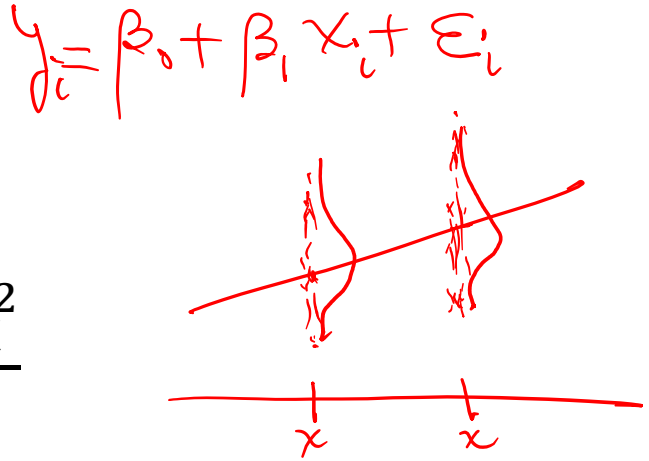
donde $S_{xy} = \text{Covarianza.muestral}$

$$= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

$$S_x^2 = S_{xx} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

$$e_i = y_i - \hat{y}_i = \ i - \text{ésimo residual}$$

Estimación de la varianza del error



Un estimador insesgado de σ^2 es:

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$$

s^2 es también llamado el error estandar residual (RSE).

Old Faithfull geyser

Para estimar la espera para la siguiente erupción se trata de ver una posible relación entre la duración de una erupción y el tiempo de espera para la erupción siguiente.

Se midieron:

duración de una erupción (X)

tiempo de espera para ver la siguiente erupción (Y)

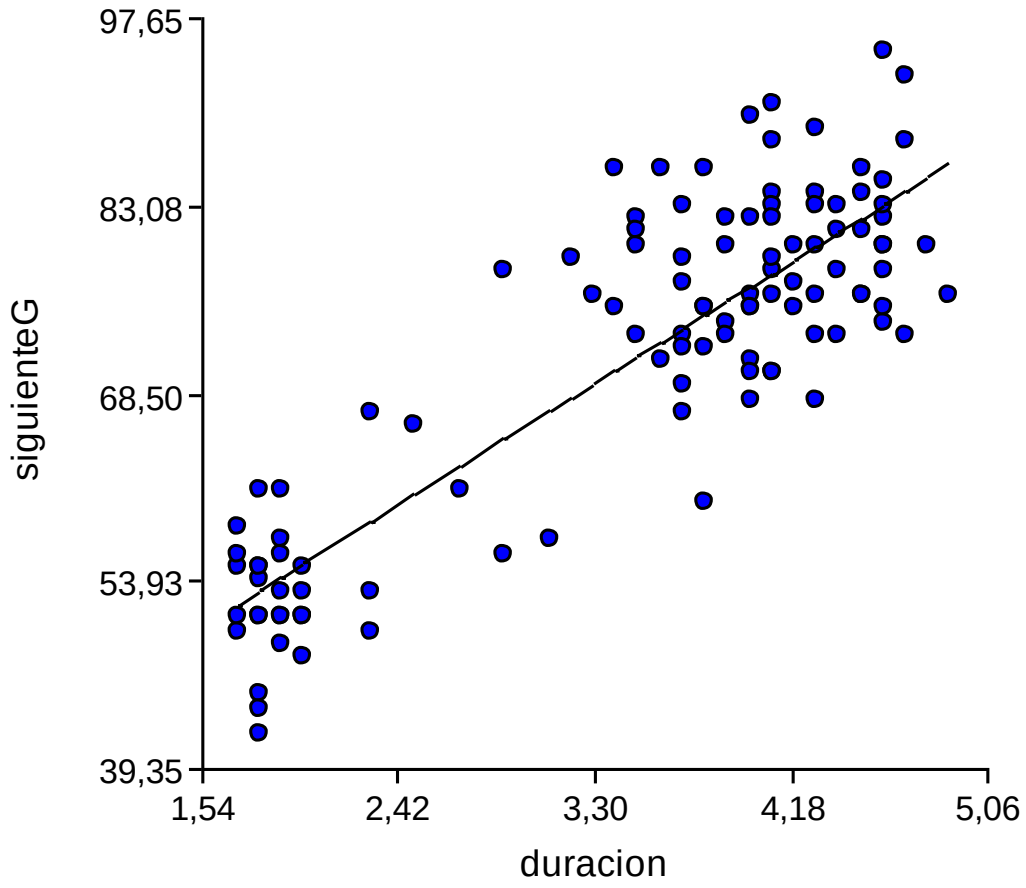
Ejemplo 1: rls para datos geyser

Coef	Est.	E.E.	LI(95%)	LS(95%)	T	p-valor
const	34,00	2,20	29,65	38,36	15,48	<0,0001
Duracion	10,69	0,61	9,49	11,90	17,62	<0,0001

La ecuación de la recta ajustada es…

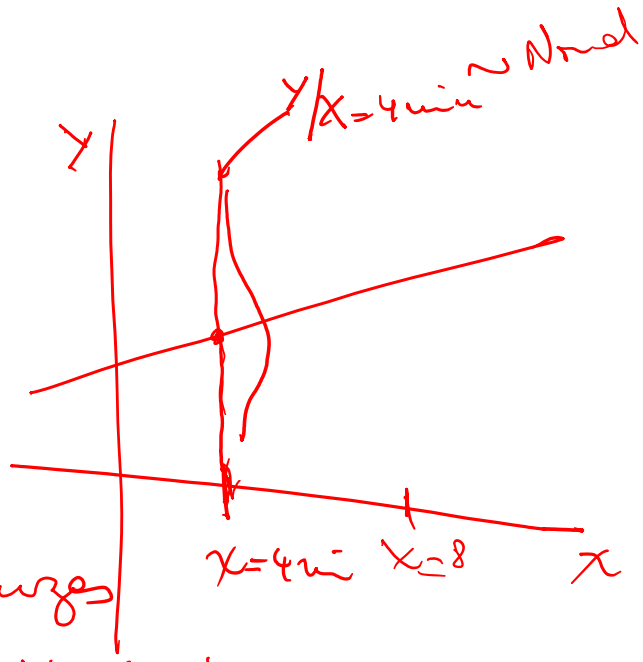
$$\hat{y} = 34 + 10.69x$$

Recta de regresión ajustada

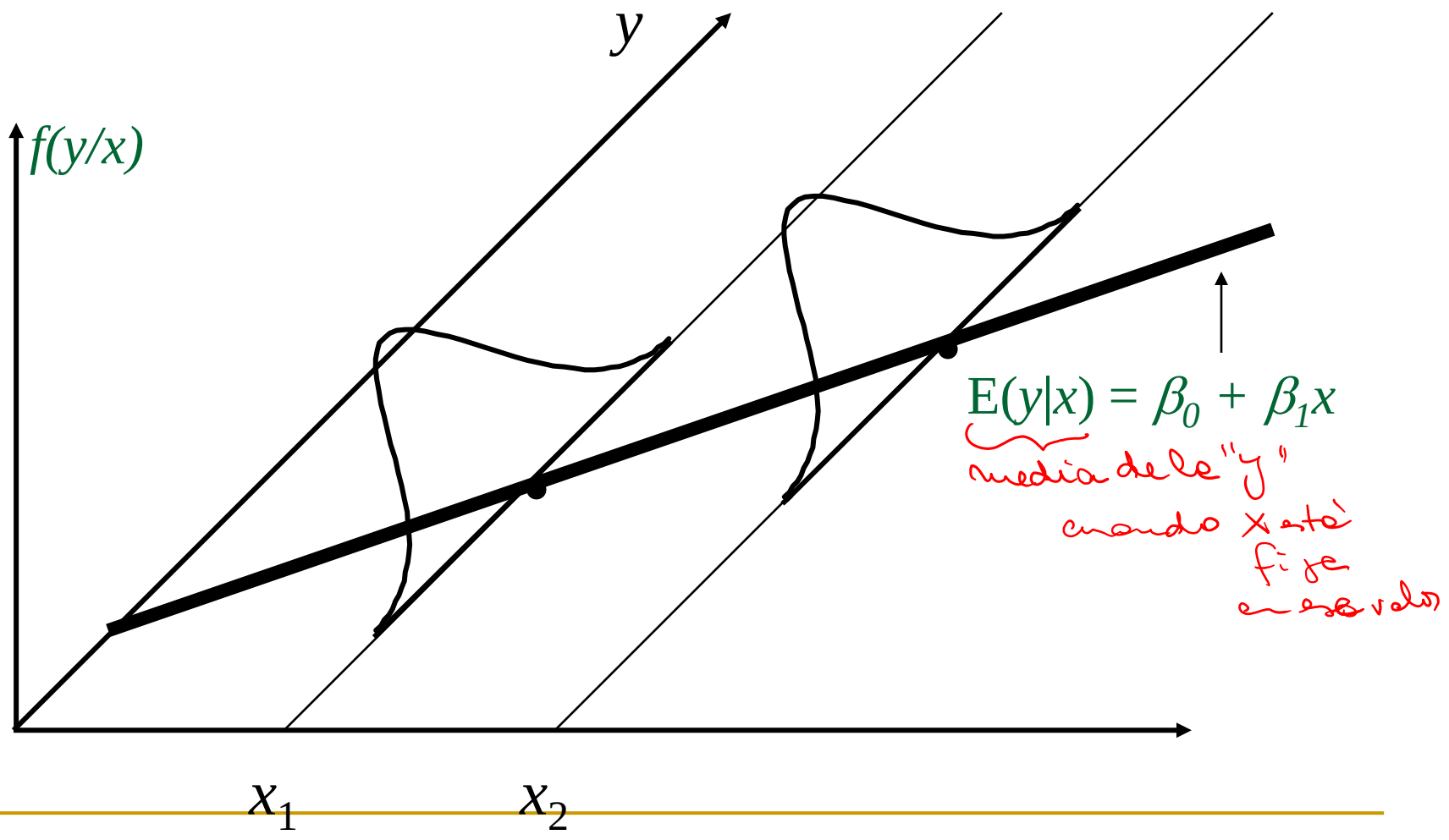


Supuestos del modelo

1. linealidad
2. regresores no estocásticos
3. $E(\varepsilon_i)=0$ y $Var(\varepsilon_i)=\sigma^2$
4. ε_i no correlacionados
5. ε_i Normales



Supuestos: la media de $y|x$ es una funcion lineal de x



Distribución de los estimadores Mínimos Cuadrados de la regresión

Los estimadores de los coeficientes son insesgados.

Es decir,

$$E(\hat{\beta}_0) = \beta_0 \text{ y } E(\hat{\beta}_1) = \beta_1$$

La varianza de $\hat{\beta}_1$ es $SE^2(\hat{\beta}1) = \frac{\sigma^2}{n.Sxx}$

y la varianza de $\hat{\beta}_0$ es $SE^2(\hat{\beta}0) = \sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{n.Sxx})$

Además, bajo el supuesto de normalidad de los errores los estimadores de los coeficientes son también Normales.

Inferencia en Regresión Lineal Simple

- Pruebas de hipótesis e intervalos de confianza acerca de los **coeficientes de regresión** del modelo de regresión poblacional.
- Intervalos de confianza para el **valor medio** de la variable de respuesta y intervalos de probabilidad para un valor **predicho**.

Para hacer inferencia acerca de los parámetros usamos que, bajo los supuestos del modelo,

$$\frac{\hat{\beta}_1 - \beta_1}{\frac{s}{\sqrt{nSxx}}} \sim t_{(n-2)} \text{ y } \frac{\hat{\beta}_0 - \beta_0}{s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{n.Sxx}}} \sim t_{(n-2)}$$

Inferencia acerca de la pendiente y el intercepto usando la prueba t.

Un intervalo de confianza para la pendiente poblacional β_1 es de la forma:

$$(\hat{\beta}_1 - t_{(n-2,\alpha/2)} \frac{s}{\sqrt{nSxx}}, \hat{\beta}_1 + t_{(n-2,\alpha/2)} \frac{s}{\sqrt{nSxx}})$$

Donde α representa el nivel de significación.

Intervalo de confianza para el intercepto β_0

Un intervalo de confianza para el intercepto β_0 de la linea de regresión poblacional es de la forma:

$$(\hat{\beta}_0 - t_{(n-2,\alpha/2)} s \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{nSxx}}, \hat{\beta}_0 + t_{(n-2,\alpha/2)} s \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{nSxx}})$$

Pruebas de hipótesis para la pendiente β_1

Usamos el mismo estadístico de prueba *t* para ensayar las hipótesis

$$Ho: \beta_1 = \beta^*$$

$$Ha: \beta_1 \neq \beta^*$$

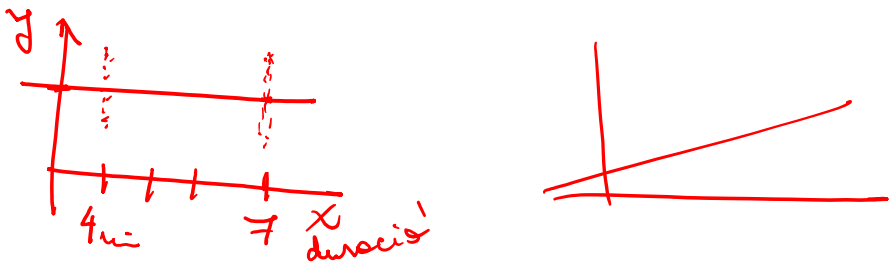
Un “p-valor” cercano a cero sugiere rechazar la hipótesis nula.

Ejemplo 1: rls para datos geyser

Coef	Est.	E.E.	LI(95%)	LS(95%)	T	p-valor
const	34,00	2,20	29,65	38,36	15,48	<0,0001
Duracion	10,69	0,61	9,49	11,90	17,62	<0,0001

$\hat{\beta}_1$

$H_0: \beta_1 = 0$ $H_1: \beta_1 \neq 0$



Prueba la hipótesis

$$\beta_1 = 0$$

En RLS:

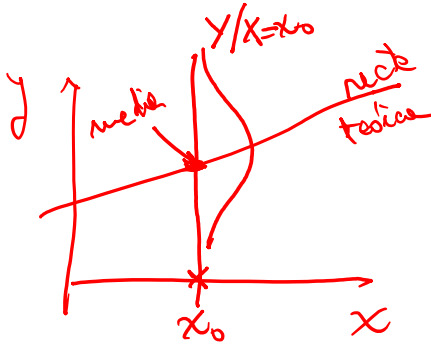
$H_0: \beta_1 = 0$ prueba la significatividad del modelo

o "¿vale la pena ajustar este modelo? para predecir Y a partir de X ??

Si NO Rechazo H_0 (p valor grande) → "no vale la pena" o sea NO es SIGNIF el modelo

En este caso, p valor chico → el MODELO es SIGNIF (Rechazo H_0)

Intervalos de Confianza para el valor medio de Y e Intervalo de Predicción



Se busca es establecer un intervalo de confianza para la media asumiendo que la relación entre X e Y es lineal.

$$\hat{Y}_o = \hat{\beta}_0 + \hat{\beta}_1 x_o$$

$1-\alpha = 0{,}95$ o $0{,}99$ etc ....

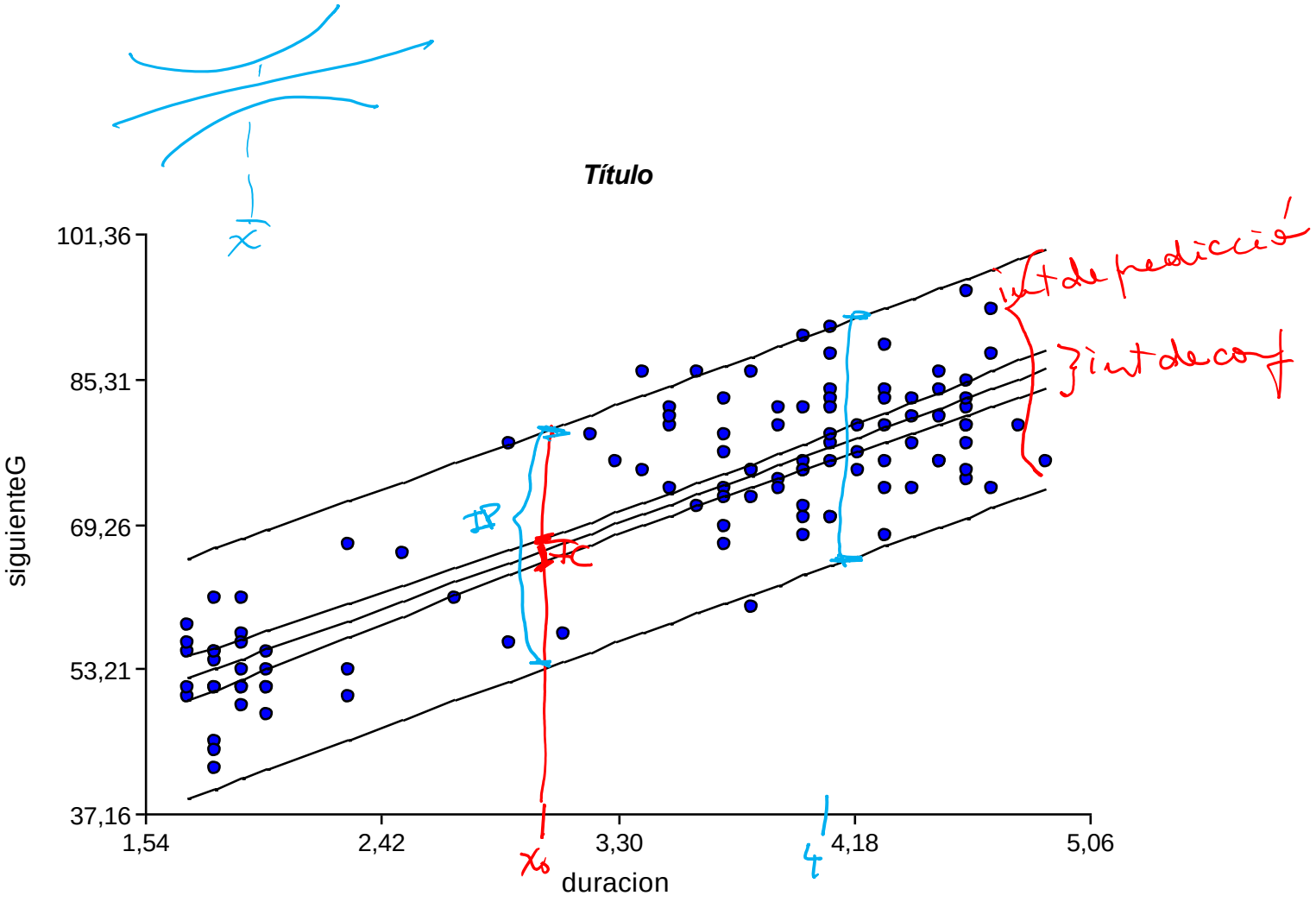
Un intervalo de confianza para el valor medio de Y dado que $X = x_0$ esta dado por:

$$\hat{Y}_0 \pm t_{(1-\alpha/2,n-2)} s \sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{nS_{xx}}}$$

Un intervalo de predicción para un valor predicho de Y dado que $X = x_0$ es de la forma:

$$\hat{Y}_0 \pm t_{(1-\alpha/2,n-2)} s \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{nS_{xx}}}$$

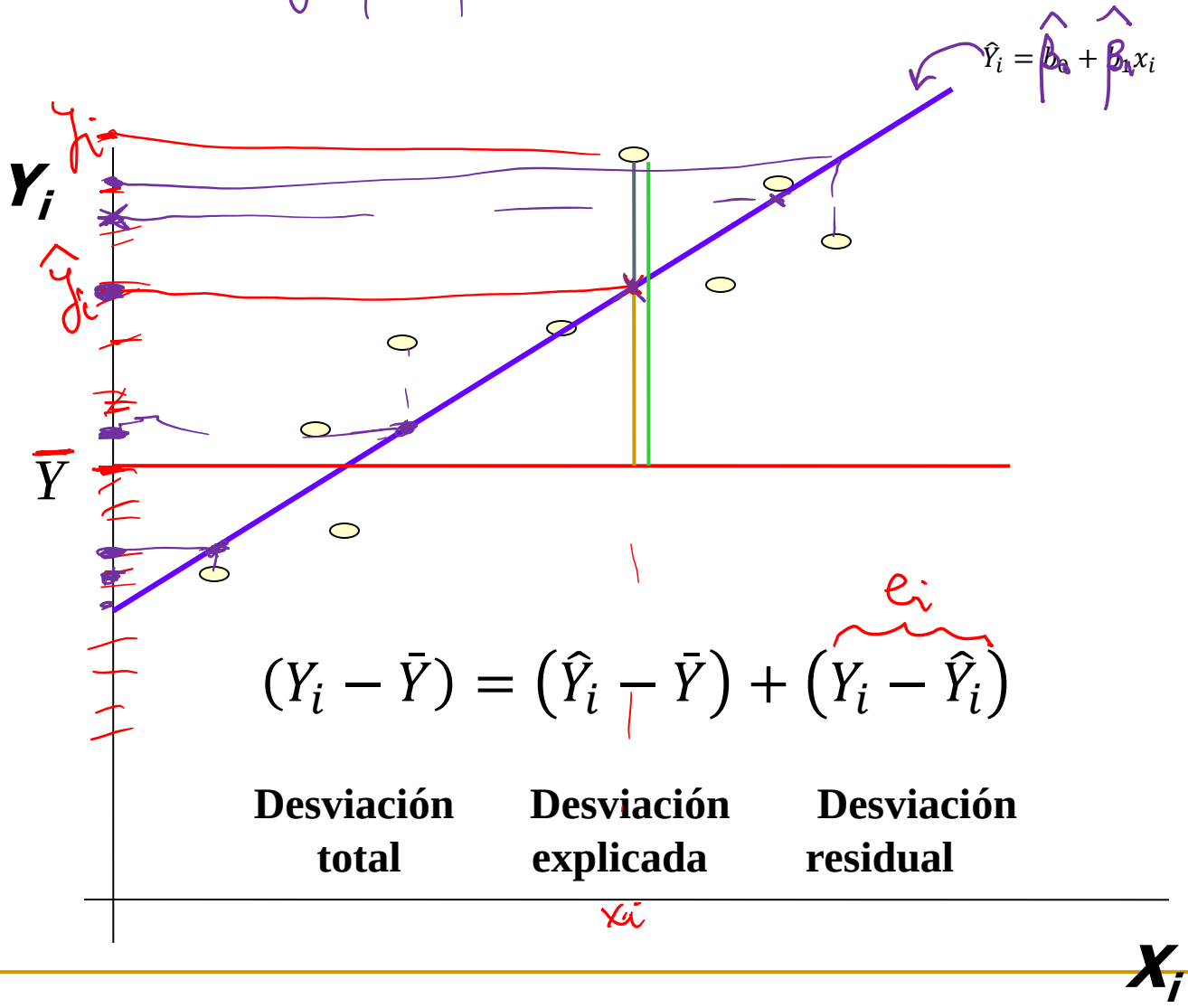
Ej Geyser: Intervalos de confianza y de predicción



Observaciones

- El Intervalo de la predicción es más amplio que el de confianza para la media.
- Ambos se usan en el rango de x's observados.
- Si x_0 está lejos de $\bar{x}$ entonces los intervalos son más amplios.

Descomposición de la suma de cuadrados total



Descomposición de la suma de cuadrados total

La desviación de un valor observado con respecto a la media se puede escribir como:

$$(y_i - \bar{y}) = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

dispersión de los 'y' obs

e_i

dispersión de los $\hat{y}_i$ (ajustados)

SST = SSErr + SSReg

El Coeficiente de Determinación R^2

Mide la relación entre X e Y (la mejor recta posible no tiene porqué ser buena)

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

Se interpreta como **la porción de variación total que está explicada por la regresión.**

Además, si *r* es el coeficiente de correlación muestral,

$$\boxed{R^2 = r^2} \quad \text{donde} \quad r = \frac{Sxy}{\sqrt{SxxSyy}}$$

El análisis de varianza para regresión lineal simple

Consiste en descomponer la variación total de la variable de respuesta en varias partes llamadas fuentes de variación.

La división de la suma de cuadrados por sus grados de libertad es llamada cuadrado medio.

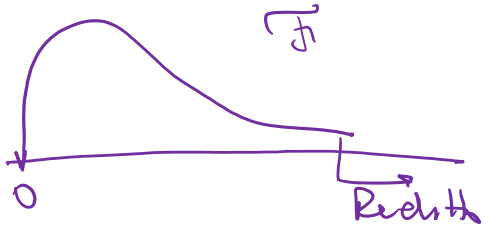
Así se tienen tres cuadrados medios:

Cuadrado Medio de Regresión=MSR=SSR/1

Cuadrado Medio del Error Residual= MSE=SSE/(n-2) =RSE

Cuadrado Medio del Total=MST=SST/(n-1)

Tabla de Análisis de Varianza



Fuente de Variación	g.l.	Sumas de Cuadrados	Cuadrados Medios	F
Debido a la Regresion	1	SSR	MSR=SSR/1	$\frac{MSR}{MSE}$
Error	n-2	SSE	MSE=SSE/(n-2)	
Total	n-1	SST		

Ho: el modelo no es signif ("vale la pena ajustar este modelo"?)

La prueba F testea la significatividad de la regresión, o sea $\beta_1 \neq 0$

Se rechazaría la hipótesis nula Ho: $\beta_1 = 0$ si el "p-valor" de la prueba de F es chico.

Esta prueba coincide con la prueba t dada antes.

Ejemplo 1: rls para datos de geyser

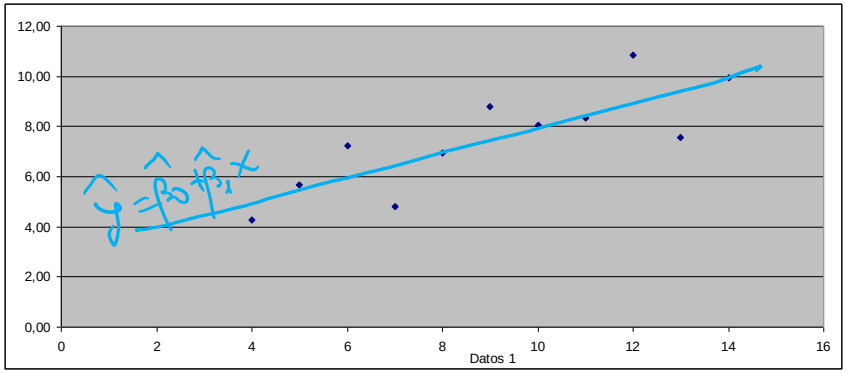
Variable	N	R^2	R^2 Aj
siguienteG	111	0,74	0,74

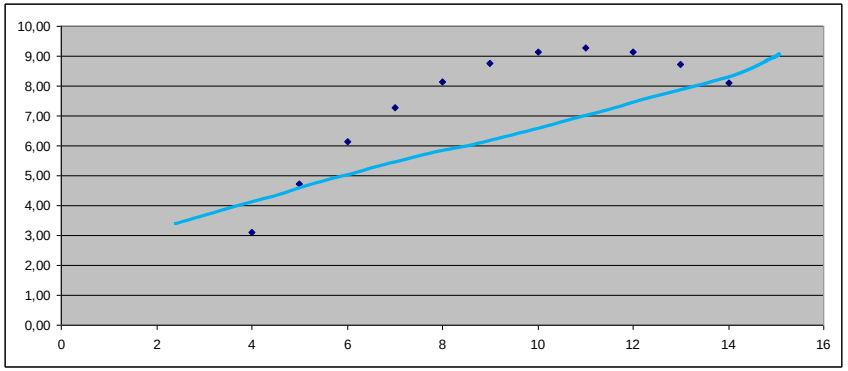
Cuadro de Análisis de la Varianza

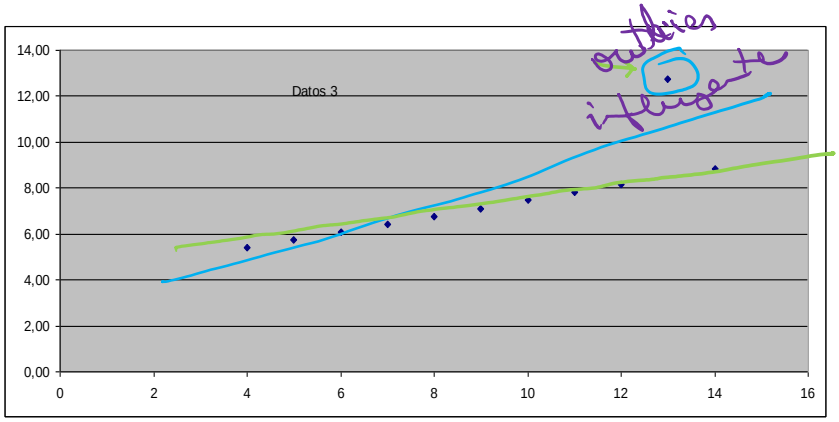
F.V.	SC	gl	CM	F	p-valor
Modelo.	13525,07	1	13525,07	310,49	<0,0001
Duracion	13525,07	1	13525,07	310,49	<0,0001
Error	4748,03	109	43,56		
Total	18273,10	110			

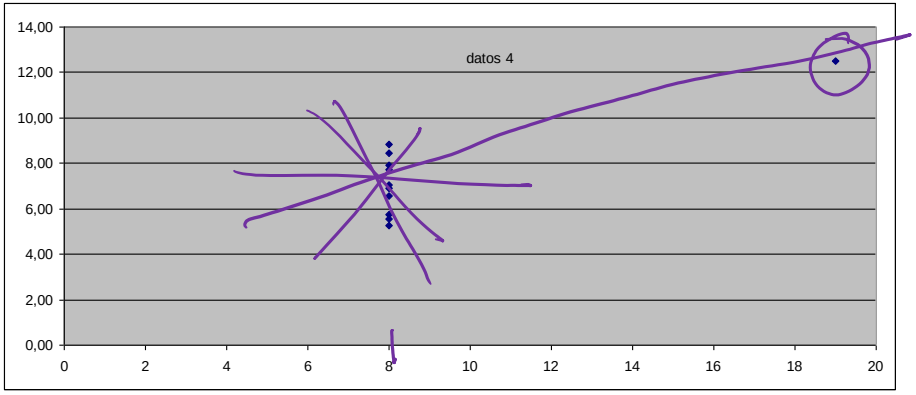
Ojo! Este coeficiente no mide bondad de ajuste! (Cuarteto de Anscombe)

American Statistician, 27 (febrero de 1973), 17-21









En todos los casos $R^2 = 0{,}66$

Análisis de Residuales

Los residuales pueden pensarse como los “errores observados” suponiendo que el modelo es correcto. Permiten evaluar si las suposiciones del modelo se cumplen y explorar el porqué de un mal ajuste del modelo. Podemos ver:

- Si la distribución de los errores es normal y sin “outliers”.
- Si la varianza de los errores es constante y si se requieren transformaciones de las variables.
- Si la relación entre las variables es efectivamente lineal o presenta algún otro patrón.

Propiedades de los residuales

Los residuales son las desviaciones de los valores observados de la variables de respuesta con respecto a la línea de regresión. Cumplen las siguientes propiedades:

a). $\sum_{i=1}^{n} e_i = 0$

b). $\sum_{i=1}^{n} e_i x_i = 0$

c). $\sum_{i=1}^{n} e_i \hat{y}_i = 0$

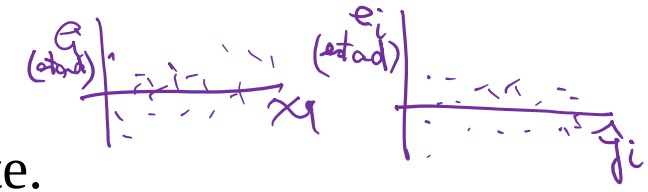
Esto permite analizar el cumplimiento de los supuestos.

Análisis de Residuales

Histograma/box-plots de Residuales: para chequear gráficamente algunos supuestos.

Plot de Normalidad de Residuales (QQ plot, PP plot): Permiten cotejar normalidad.

Plot de Residuales versus los valores predichos : Se usa para detectar si hay datos anormales, cuando hay datos que caen bastantes alejados, tanto en el sentido vertical como horizontal. Si se usan residuales estandarizados, entonces un dato con residual más allá de 2 ó -2 es considerado un "outlier" en el sentido vertical**.**

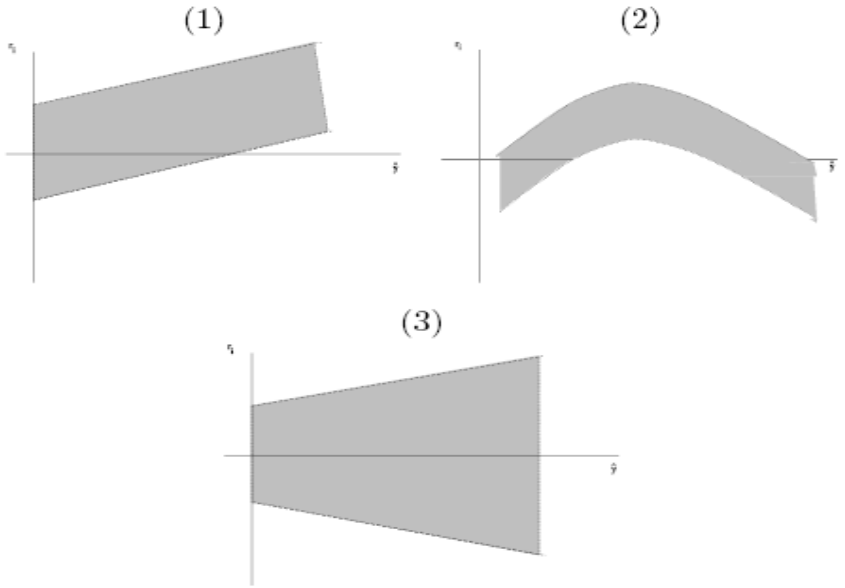


También sirve para ver si la varianza es constante.

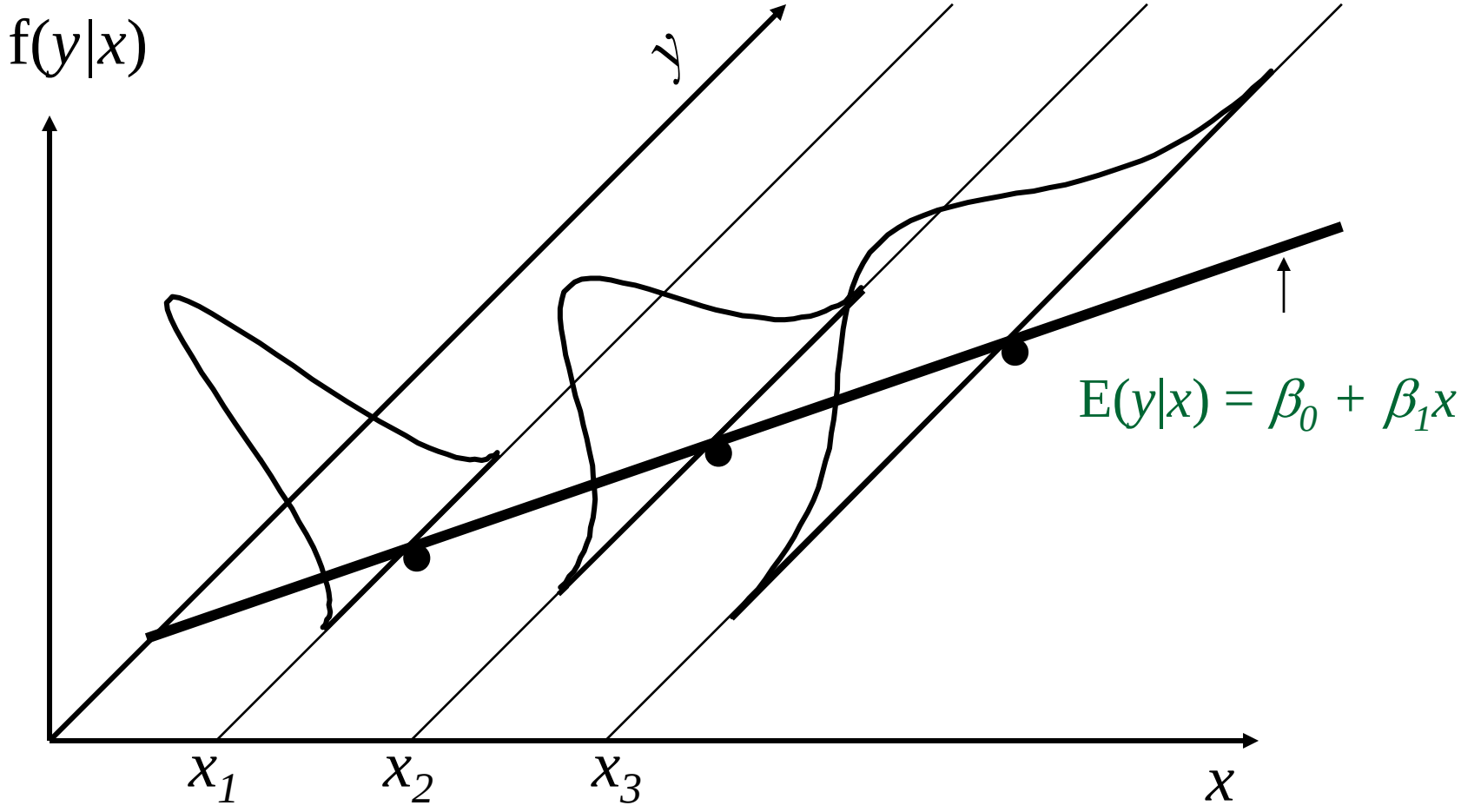
Plot de Residuales versus x: se usa con igual criterio que con los predichos.

Plot de residuales para chequear linealidad y atípicos

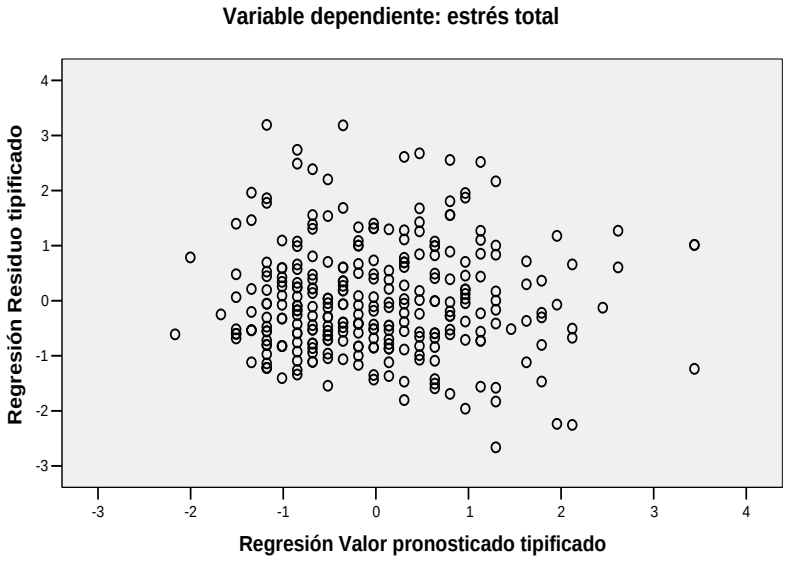
Graficamos residuales versus predichos para chequear estos supuestos. En los siguientes gráficos se ven indicios de (1) modelo no adecuado; (2) falta de linealidad; (3) no homogeneidad de varianzas (heterocedasticidad).

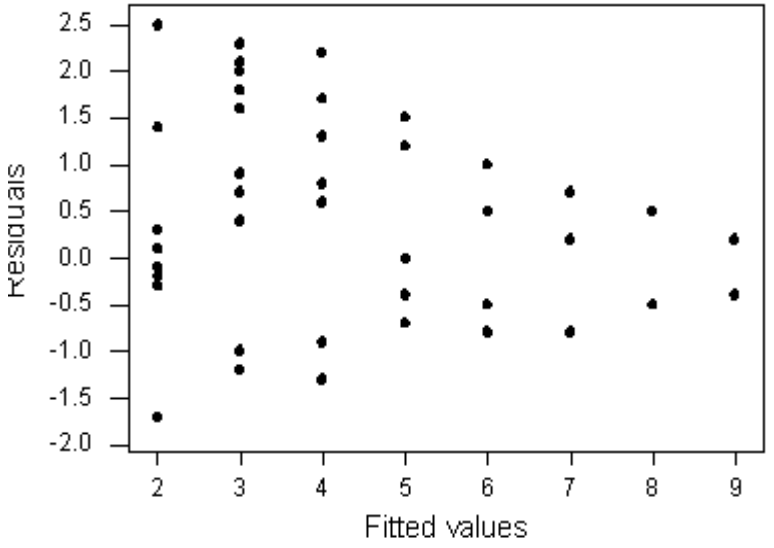


Qué significa heteroscedasticidad?



Gráficos de residuales vs predichos





El primer gráfico muestra un comportamiento adecuado de los residuos. El segundo muestra que no hay homogeneidad de varianzas.

Remedios cuando la varianza σ^2 no es constante

- Usar mínimos cuadrados ponderados donde los pesos que se usan son hallados en base a los datos tomados.

- Transformar la variable de respuesta Y usando transformaciones que estabilizan la varianza

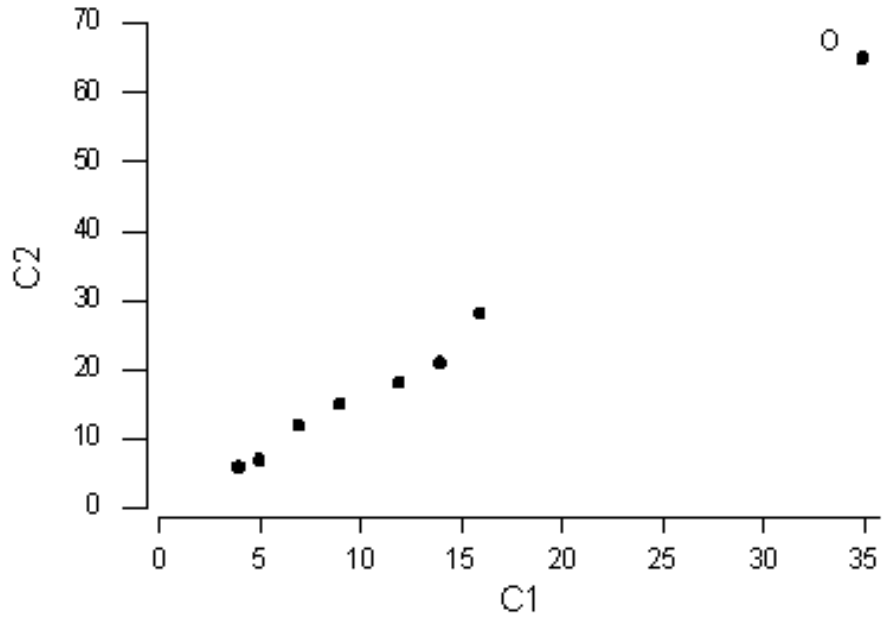
Outliers y puntos influyentes

Una observación es considerada un **outlier** si está alejado de la mayoría de los datos (leverage alto) sea en la dirección de Y o en la de alguno de los regresores ó en ambos.

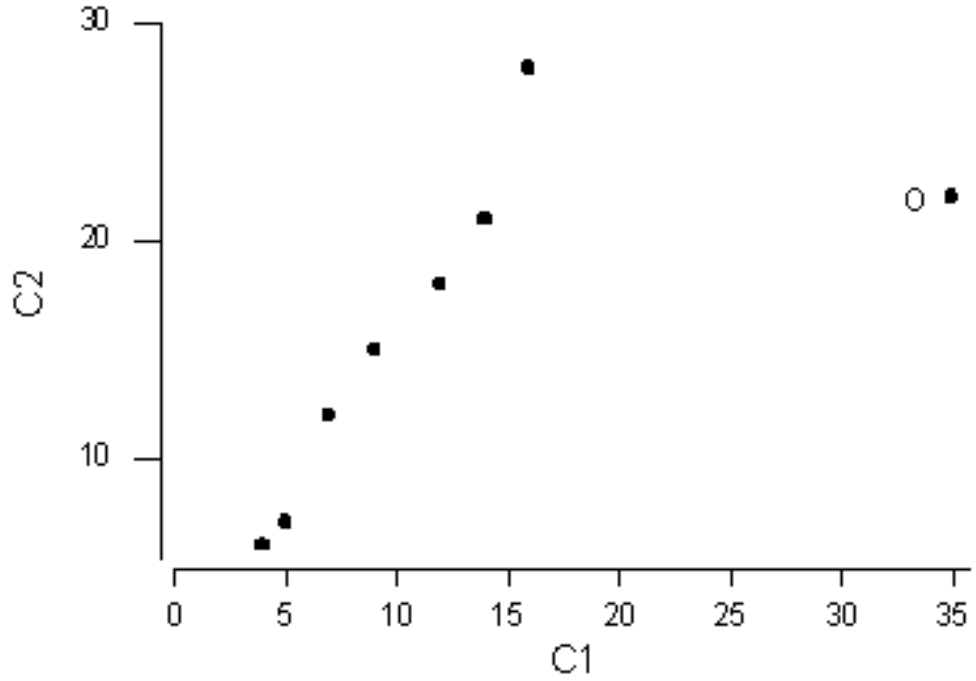
Diferenciamos outliers en Y (se los ve en los gráficos de residuales) de outliers en x's.

Una observación es considerado un **valor influyente** si su presencia afecta de manera importante el comportamiento del modelo. Por ejemplo en el caso de regresión simple remover un valor influencial cambiaría mucho el valor de la pendiente.

Ejemplo de una observación que es "outlier" y punto leverage alto pero que no es influencial.



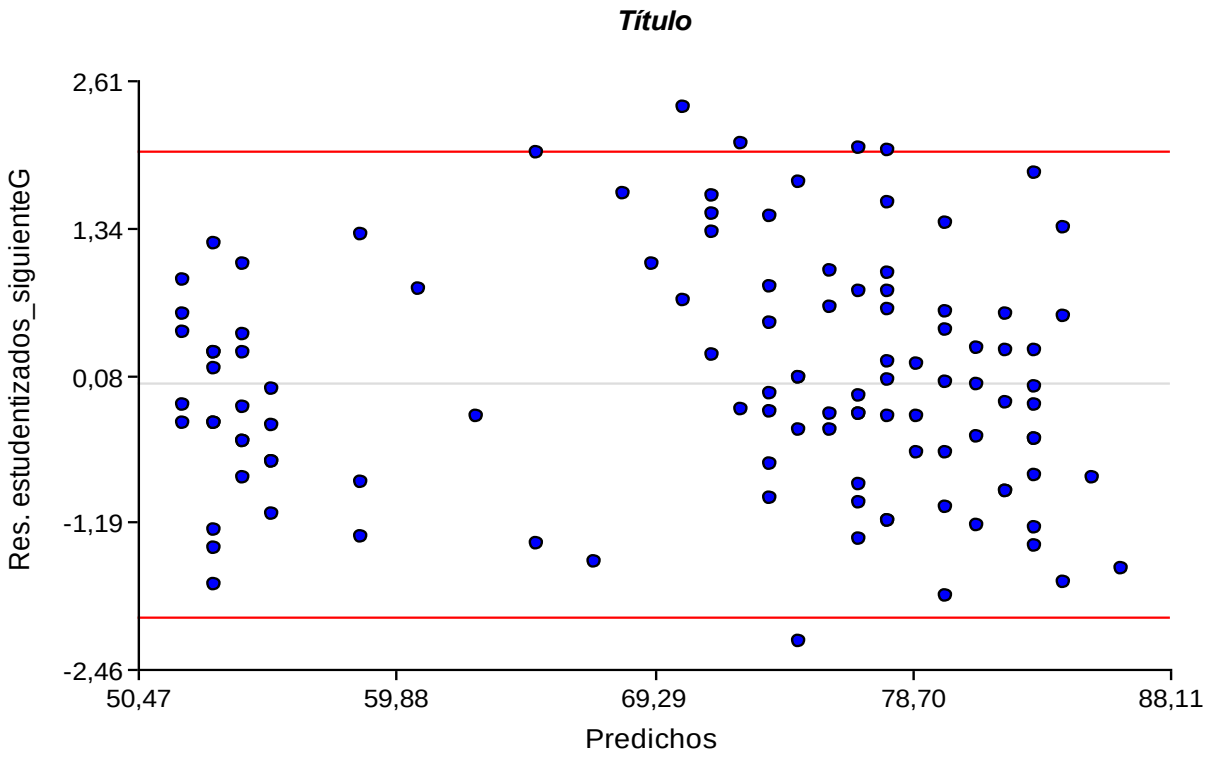
Ejemplo de una observación que es punto de leverage alto y que también es influencial.



Este punto tendrá un gran efecto sobre el R^2 y da un cambio drástico en la pendiente.

Ejemplo: rls para geyser

Veamos los supuestos en el gráfico de residuales:



Parece adecuado el supuesto de igual varianza en cada nivel (homogeneidad), así como la linealidad.